Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Computer Science 17 (2013) 797 – 802
Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

The Research of Weighted Community Partition based on SimHash

Li Yang[a,b,c,] **, Sha Ying[c], Shan Jixi[c], Xu Kai[c]

[a]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190*
[b]*Graduate School of Chinese Academy of Sciences, Beijing, 100049*
[c]*National Engineering Laboratory for Information Security Technologies, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093*

Abstract

The methods of community partition in nowadays mainly focus on using the topological links, and consider little of the content-based information between users. In the paper we analyze the content similarity between users by the SimHash method, and compute the content-based weight of edges so as to attain a more reasonable community partition. The dataset adopt the real data from twitter for testing, which verify the effectiveness of the proposed method.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: Social Network, Community Partition, Content-based Weight, SimHash, Information Entropy;

1. Introduction

The important characteristic of social network is the community structure, but the current methods mainly focus on the topological links, the research on the content is very less. The reason is that we can't effectively utilize content-based information from the abundant short text. So in the paper we hope to make full use of content-based information between users based on the topological links, so as to find the more reasonable communities.

Nowadays, the research of social network mainly focus on these: (1)Graph Partition Method[1-4]. However graph partition need to confirm the amounts of group and even value in advance, we cannot learn the value in advance, therefore the method cannot be applied for community detection directly. (2) Divided Method[5]. Specifically speaking, it removes the largest side of the interface (Betweenness) from the network through the iteration and divides the whole network into many communities. But in the case of unknown number of

* Corresponding author: Li Yang.
E-mail address: leoncs2000@gmail.com.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.102

community, the algorithm can't determine the right number of iterations. (3) Agglomeration Method[6-8]. For the each random node i, j, when the node i join the community containing node j, we compute the corresponding increment of modularity: ΔQ, and decide the node i whether join or not according to the judgement of ΔQ. The merger process repeat until the entire network can't appear merger phenomenon. (4) Overlapping Community Detection[9-11]. In the actual network, some nodes are often shared by many communities, it means that there is overlap between communities.

In the research of this paper, based on the numerous short text in the social network(mainly focus twitter in this paper), we plan to adopt the SimHash to deal with the content, quickly computing the content-based weight between users relying on the topology, and then resort to the agglomeration method to partition the more reasonable communities.

The structure of this paper is as follow: The first part is the introduction, the second part is the relative research, the third part is the weighted community partition based on SimHash, the fourth part is the result and analysis of experiment, and the last one is the summary.

2. **The Relative Research of Content Computation**

Currently, there are some methods in the fields of the content similarity computation.(1) Content-based Dictionary[12-17]. The method based on content-based dictionary is the most widely used method. However there are some disadvantage, one is that the dictionary is impossibility complete, the other is that the polysemia of words will reduce the accuracy of similarity computation. (2) Based on Large-scale corpus[18-20]. By means of extracted words in every document we can construct a matrix of words-document, the one deficiency is the newest document perhaps contain some novel words, another is that the matrix based on these short text will is very sparse, which will influence the accuracy of similarity computation.(3) Content-based Feature[21-23]. This method attempt to indicate the text by using the predefined features, the difficulty of the method is how to define the effectual feature and automatically get the value of features. However these methods can't be used to deal with the short text in the social network.

3. Weighted Community Partition based on SimHash

3.1.The definition of problem

- The current research on community partition mainly focus on the topological links, but the two users contain abundant content-based information, so how to construct the similarity analysis between content is the first problem that we need to consider.
- Meanwhile, the content in the social network are full of short texts, the conventional method (such as in blog or bbs) can't be applied to the short text, so how to deal with the short text is the second problem we need to consider.
- Finally, how to assess the quality of community partition based content-based weight is the third problem that we need to consider.

3.2.Characteristics analysis of tweet in twitter

Twitter[24] is one of the most representative platforms of the micro-blog. One message on twitter we call it a tweet. On Twitter people can send tweet. People can not only reply to the tweet that they are interested, but also they can forward the tweet, that is to say the same tweet was sent again by the other people who are interested. The benefit is, if someone send a tweet and the tweet will become a news in the future, then the

followers will forward the tweet and the followers' followers will also forward it, the last the tweet perhaps will be forward millions times in a few minutes.

The forwarder can copy messages that he was interested and sending out with a little modification. The modification makes the tweet is different from the origin tweet, but the means are almost the same. So, the main idea of this part is to calculating the rate of the similar tweets and constructing the content-based weight of the tweet contents that they have sent. This is the first problem that we need to solve.

3.3. SimHash method

Twitter has a large of short texts are forwarded and similarity. In the paper we introduce the SimHash[25] to extract the same or the similar tweets, as shown in the fig 1. The core concept of the SimHash is reducing the dimensionality of the text vector, converting a high dimensionality text vector to a integer which is 64-bit in our paper and the 64-bit integer we call it fingerprint, in our paper each fingerprint represent a tweet and the fingerprint similarity represents the tweet similarity, the number of the different binary-bit between two fingerprints is called hamming distance.

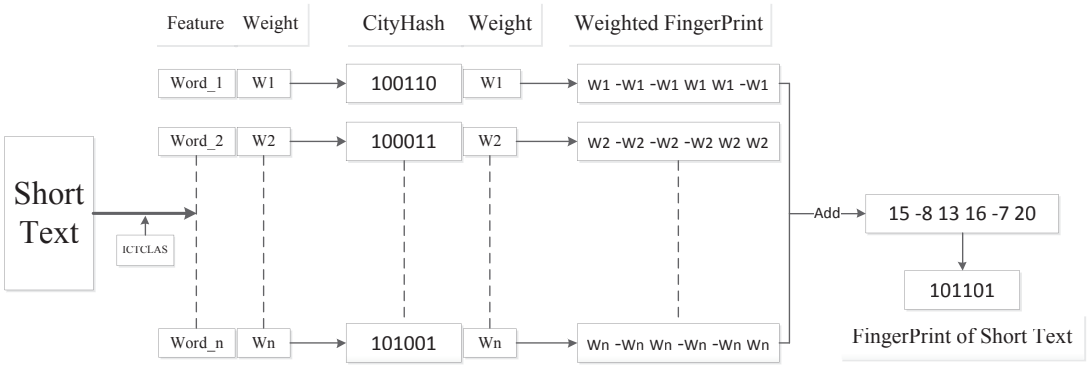

Fig.1.The computation process of SimHash

SimHash needs the characteristic set of the document, characteristic set contains the words set and the weight of each word, for getting the words set, we use the ictclas[26]word segmentation system made by Chinese academy of sciences. In view of the short text character, we put all the words as the character word after the word duplicate removal, give each word with one weight, then using cityhash[27] to calculate the signature of each word. This is the second problem that we need to solve.

3.4.Weighted community partition based on SimHash

3.4.1.The Computation of Content-based Weight

Firstly, based on the structure network, for the each edge E(M,N are the two users),extracting all the relative contents(such as post, tweet, comment),and then carrying out the similarity computation of content between M and N based on the SimHash method, finally we can get the content-based weight between M and N(M_c and N_c indicate the relative content respectively),the formula is:

$$SIM_{semantics}(M,N) = \left| \frac{M_c \bigcap N_c}{M_c \bigcup N_c} \right| \quad (1)$$

3.4.2.Main steps

Table 1. Steps of community detection based on SimHash

Input: The Graph of Twitter network : G = (V,E)，V indicates the whole users set, E indicates the whole edge set(such as following, tweet, comment), C_i indicate the whole content of user N_i;
Output: Content-based Communities C={C_1,C_2,…,C_k}， C_j(1<=j<=k) indicates the sub-community;
1. Integrate the network structure of all the user;
2. Deal with the content of every user, such as word segmentation, remove stop words, remove sign;
3. As for the edge between two users, compute the content-based weight based on the SimHash;
4. Based on the weighted network structure, we begin to partition community so as to discover the most reasonable communities.

4. The Result and Analysis of Experiment

4.1 Experiment environment

4.1.1.Running environment

- CPU: Quad-Core of AMD.
- OS: Win 7 32-bit.
- DataSet: Twitter data acquired from internet, the scale of dataset are 10 million, 20 million, 50 million, 100million respectively.

4.1.2 Comparative methods

Adopt the following three methods to compute the content-based weight between users, compare the influence on the weighted community partition.

- Vector Space Model: By accumulating the all content of the user, we can attain the vector space based on the content's feature, then compute the similarity based on the cosine value between users.
- Levenshtein Distance: Compute the content-based weight between users based on the proportion to high similarity of content.
- SimHash: The method is proposed above in this paper.

4.1.3.Assessment index

- Running Time: The main work of this part is to assess the running time of different methods computing weight, which can influence the further community partition.
- Information Entropy of Community: Definition 1 The number of the whole user of the network is n, the number of sub-communities is m, here, i=1,2,…,n indicates the identifier of user, j=1,2,…,m indicates the identifier of edge; In the every sub-community, there are z different weight of edge(such as d_1,d_2,…,d_z), therefore we define the information entropy of sub-community is S_j, and the information entropy of entire community is S, which are as follows:

$$S_j = -\sum_{x=1}^{z} P(d_x)\log P(d_x) \qquad (2)$$

$$S = \frac{1}{m} * \sum_{j=1}^{m} S_j \quad (3)$$

$P(d_x)$is the probability of dx, that is $P(d_x) = k/ N_{Cx}$, k is amount of the content-based weight dx, N_{Cx} is amount of edge in the sub-community of S_j. The smaller the content-based entropy of community S is, the more consistent the users of the community are, that is the users in the community have common interest and the structure of communities are more reasonable, and vice versa.

4.2. Analysis of results

4.2.1. Comparative analysis based on time

Table 2. Running time of content-based weight based on different scales of data(Second)

Method	1W	2W	5W	10W
Vector Space Model	11332	11427	14702	15156
Levenshtein Distance	1378	4461	58209	79027
SimHash	33	69	160	557

Analysis of efficiency: as shown in table 2, when the scale of dataset is small, the running time of Levenshtein distance is in medium, but with the increase of scale of dataset, the running time explosively grow, because the time complexity of Levenshtein distance is $O(n^2)$.

The base running time of vector space model is large, which need a series of proceed, such as word segmentation, vectorization and computation of TFIDF, but the whole running time grow slowly with the increase of scale of dataset.

As for the method of SimHash, that is the proposed method in this paper, we can improve the matching speed of Hash mapping, accelerate the calculation speed, finally attain impressive effect, which verify the efficiency of SimHash in fields of short text.

4.2.2.Comparative analysis based on information entropy

Table 3. Information entropy of community based on different scales of data

Method	1W	2W	5W	10W
Vector Space Model	0.0498618	0.0678264	0.134733	0.338977
Levenshtein Distance	0.0510738	0.082282	0.0556155	0.0506236
SimHash	0.010711	0.00711932	0.035672	0.0410565

Analysis of information entropy: as shown in table 3, the content-based entropy of community based on vector space model is maximum, the Levenshtein distance takes second place, the content-based entropy based on SimHash is minimum, therefore we can see that we can attain the most reasonable commnnities based on SimHash.

5.Summary

By analyzing and researching of the content-based weight between users in twitter, we propose the SimHash method to compute the similarity of short text quickly, and then we can partition community based on content-based weight. Comparing with the methods of levenshtein distance and vector space model, we can demonstrate the advantage of SimHash which can quickly compute the similarity of content, further we can attain more reasonable communities.

Acknowledgements

This research was supported by The National Natural Science Foundation of China (61070184), the "Strategic Priority Research Program" of the Chinese Academy of Sciences，Grant No.XDA06030200, and National Key Technology R&D Program(No:2012BAH46B03).

References

[1] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs[J].Bell System Tech. J., 1970,49(2):291-307.
[2] E. R. Barnes. An algorithm for partitioning the nodes of a graph[J]. SIAM J. Alg.Disc. Math., 1982,3(4):541-550.
[3] G.W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities[C]. In SIGKDD'00: Proceedings of the sixth ACM Int Conf on Knowledge Discovery and Data Mining, 2000: 150-160.
[4] G. W. Flake, S. Lawrence, C. L. Giles,et al. Self-organization and identification of Web communities[C]. IEEE Computer, 2002,35:66-71.
[5] Girvan M, Newman MEJ. Community structure in social and biological networks [C]. Proc. Natl. Acad. Sci, 2002, 9(12):7821 - 7826.
[6] New man M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E,2004,69(2):026113.
[7]Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Phys Rev E,2004,70(6):066111.
[8] Blondel V D, Guillaume J L, Lambiotte R,et al. Fast unfolding of community hierarchies in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment,2008,10:10008.
[9] SHEN Hua-wei, CHENG Xue-qi, CAI Kai,et al. Detect overlapping and hierarchical community structure in networks[J].Physica A, 2009, 388(8): 1706 -1712.
[10] G. Palla, I. Derenyi, I. Farkas,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 435:814-818, 2005.
[11]. G. Palla, A.-L. Barabasi and T. Vicsek. Quantifying social group evolution[J]. Nature, 446:664-667, 2007.
[12] Fellbaum C.1998.WordNet:An Electronic Lexical Database[M],The MIT Press.
[13] HowNet[EB/OL].1999.http://www.keenage.com/.
[14] Leacock C and Chodorow M.1998.Combining local context and WordNet sense similarity for word sense identification[M].WordNet,an Electronic Lexical Database,The MIT Press.
[15] Wu Z and Palmer M.1994. Verb Semantics and Lexical Selection[C].In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,New Mexico State University,Las Cruces,New Mexico,US,133-138.
[16] Mihalcea R,Corley C and Strapparava C.2006.Corpus-based and Knowledge-based Measures of Text Semantic Similarity[C].Proceedings of the 21st National Conference on Artificial Intelligence and The Eighteenth Innovative Applications of Artificial Intelligence Conference.
[17] Islam A and Inkpen D.2008.Semantic Text Similarity using Corpus-based Word Similarity and String Similarity[J].ACM Transactions on Knowledge Discovery from Data,2(2).
[18] Landauer T,Foltz P and Laham D.1998.An Introduction to Latent Semantic Analysis[J].Discourse Processes,25(2-3):259-284.
[19] Foltz P,Kintsch W and Landauer T.1998.The Measurement of Textual Conherence with Latent Semantic Analysis[J].Discourse Processes,25(2-3):285-307.
[20] Burgess C,Livesay K and Lund K.1998.Explorations in Context Space:Words,Sentences,Discourse[J].Discourse Processes,23(2-3):211-257.
[21] McClelland J and Kawamoto A.1986.Mechanisms of Sentence Processing:Assigning Roles to Constituents[M].Parallel distributed processing:explorations in the microstructure of cognition,psychological and biological models,MIT Press,Cambridge,MA,272-325.
[22] Hatzivassiloglou V,Klavans J and Eskin E.1999a.Detecting Text Similarity over Short Passages:Exploring Linguistic Feature Combinations via Machine Learning[C].Proceedings of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora.
[23] Hatzivassiloglou V,Klavans J and Eskin E.1999b.Detecting Similarity by Applying Learning over Indicators[C].Proceedings of 37th Annual Meeting of the Association for Computational Linguistics.
[24] twitter［EB/OL］. 2012. URL：http://www.twitter.com.
[25] G. S. Manku, A. Jain, A. D. Sarma. Detecting near-duplicates for web crawling [C]. Proceedings of the 16th International World Wide Web Conference. Banff, Alberta, Canada. May , 2007.
[26] Ictclass［EB/OL］. 2012. URL：http://ictclas.org/.
[27] Cityhash［EB/OL］. 2012. URL：http://code.google.com/p/cityhash/.